

COMPARISON OF MEANS FROM POPULATIONS WITH UNEQUAL VARIANCES

Barbara A. Grimes and Walter T. Federer

School of Public Health, University of Texas, Houston, Texas
and

Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, New York, 14853

BU-762-M*

Revised August 1982

Abstract

The problem of comparing two independent sample means arising from populations with unequal variances has been considered for many years. Historically, it has become known as the Behrens-Fisher problem. In this paper we trace the historical development of work on the problem and consider a generalization of the Behrens-Fisher problem to contrasts of more than two population means. We are concerned with testing the null hypothesis $\sum_{i=1}^k c_i \mu_i = 0$ where μ_i is the mean of the i^{th} population, k is the number of populations in the contrast, and the c_i 's are real numbers such that $\sum_{i=1}^k c_i = 0$. A test of this null hypothesis is based on the sample statistic $d = \sum_{i=1}^k c_i \bar{X}_i / \sqrt{\sum c_i^2 s_i^2 / n_i}$, where \bar{X}_i , s_i^2 , and n_i are the mean, variance and sample size in the i^{th} sample.

To obtain our generalizations, we propose four tests to compare linear combinations of k means from populations with unequal variances. Two of the tests are based on the work of W. G. Cochran and two are based on the work of B. L. Welch. We evaluated the size and power of each test using numerical analytic techniques. Invariance to change in scale and asymptotic normality properties of the four tests were studied. Small sample properties were studied via simulation. The cases of three, four, and eight population means with various sample sizes and configurations of sample sizes were investigated. The differences in population variances for four treatment means ranged from 1:4 to 1:8 with two

* In the Biometrics Unit Series, Cornell University, Ithaca, New York, 14853.

different configurations of the 1:8 ratio. Different contrasts were also investigated. The purpose of these variations in structure of a contrast, of variance configurations, and of sample size configurations was to determine their effect on size and power of the four tests.

1. INTRODUCTION AND A REVIEW OF THE LITERATURE

The problem of comparing independent sample means arising from two populations with unequal variances has been studied for many years and there is a sizable literature. Historically, this problem has come to be known as the Behrens-Fisher problem. In the present paper, we are interested in individual specified contrasts involving more than two means. We propose four tests, two based on W. G. Cochran's work, for comparisons of k means for the unequal variance situation. We evaluate the size and power of each test using numerical analytic techniques.

Before proceeding with a literature review, we will establish some notation and definitions. Let π_i represent a normal population with mean μ_i and variance σ_i^2 for $i = 1, \dots, v$. From each population we draw an independent random sample. Let X_{ij} represent the j^{th} observation from the i^{th} sample, $j = 1, \dots, n_i$ and $i = 1, \dots, v$. Thus, $X_{ij} \sim N(\mu_i, \sigma_i^2)$. The usual unbiased estimators of μ_i and σ_i^2 are $\bar{X}_i = \sum_{j=1}^{n_i} X_{ij} / n_i$ and $s_i^2 = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 / (n_i - 1)$. Denote $n_i - 1$ by f_i and the tabulated value of Student's t distribution for h degrees of freedom at the α percentage point by $t_{\alpha}(h)$. The range of summation will be omitted where it is clear from its context. We are concerned with a linear contrast of k of the v sample means, say $\sum_{i=1}^k c_i \bar{X}_i$, where the c_i 's are real numbers such that $\sum_{i=1}^k c_i = 0$. We want to test the null hypothesis: $\sum_{i=1}^k c_i \mu_i = 0$. The sample statistic which we will consider is

$$d = \sum c_i \bar{X}_i / \left(\sqrt{\sum c_i^2 s_i^2 / n_i} \right) .$$

In the two-sample case, d reduces to

$$d = (\bar{X}_1 - \bar{X}_2) / \sqrt{s_1^2/n_1 + s_2^2/n_2} .$$

The comparison procedures we will be discussing are approximate, i.e., the supposed significance level or nominal significance level at which a test is performed is not the actual significance level of the test. We will speak of a test as being conservative if its actual significance level is less than the nominal level. We will call the test liberal if the actual significance level exceeds the nominal level.

In the following the history of the problem of comparison of means from populations with unequal variances is reviewed with attention restricted to those procedures designed for single-stage sampling schemes. Note that unless explicitly mentioned to the contrary, the solutions cited in the literature are for the two-sample case. The time line presented in Figure 1 gives a historical perspective of the Behrens-Fisher problem. The authors whose names are underlined proposed solutions to the problem and in some cases compared their solutions with other candidates. The other authors evaluated previously proposed solutions. Each solution is based on the idea of comparing d with a critical value given by an expression involving the sample variances and sample sizes.

The earliest proposed solution appeared in a paper by Behrens (1929). Fisher (1935;1941), while acknowledging some errors in Behrens' work, claimed to justify Behrens' solution by the use of fiducial inference. This solution (BF test) consists of comparing the value of the sample statistic d with a critical value given by an asymptotic series involving the sample variances and sample sizes. Sukhatme (1938) published tables of the 5% and 1% significance levels of the BF test. In the 1940's, W. G. Cochran produced an empirical approximation based on the Student's t -table by an inspection of Sukhatme's tables.

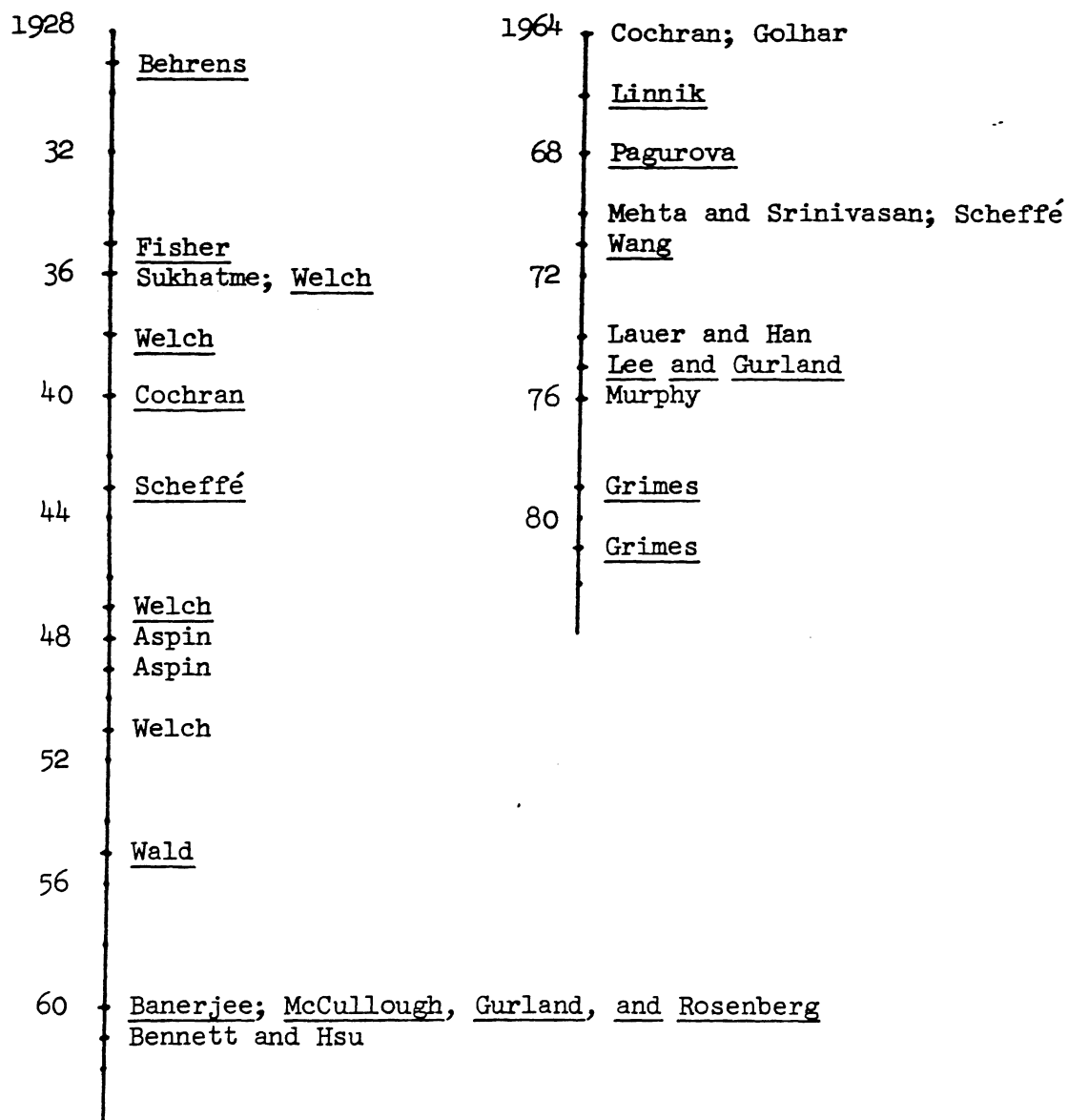


Figure 1. History of Behrens-Fisher Problem

Underlined - Proposed new solution Not underlined - evaluated solutions

His approximation was passed around by word of mouth and subsequently incorporated into a number of textbooks [see, e.g., Snedecor (1946, pages 83-84) and Federer (1955, pages 94-95)]. This led Cochran in 1964 to publish an account of the accuracy of his approximation for the two-sample problem. For $k = 2$, Cochran (1964) suggested that d could be compared to an approximate critical value t'_α , which is obtained as a weighted sum of Student's t values, namely, $t'_\alpha = [w_1 t_\alpha(f_1) + w_2 t_\alpha(f_2)] / (w_1 + w_2)$ where $w_i = s_i^2 / n_i$. We discuss his comments concerning the accuracy of the approximation later.

In a series of papers, Welch (1938; 1947; 1951) disputed Fisher's use of fiducial inference and rejected the claim that Behrens' solution had been justified; he presented an approximate solution to the BF problem (to be discussed later) and published an asymptotic solution which was further studied by Aspin (1948; 1949). Scheffé (1943) obtained a statistic for the BF problem by minimizing the length of the confidence interval for the difference of the means of two normal populations with unequal variances based on the Student's t distribution. The calculation of his confidence interval involved taking differences between sample values of the two samples. The sample sizes could be equal or unequal. The pairing of the sample values for subtraction was done randomly. The length of his confidence interval depended upon the arrangement of the sample values after random pairing.

The next proposed solution was by Wald (1955). He wanted to construct a test where the actual level of significance would not vary as the unknown population variances varied. For the equal sample sizes case, he produced a test where the actual significance level varied only slightly as the variances varied. Pagurova (1968) generalized this test to the case of two samples of unequal sizes. Pagurova's test consists of comparing the value of d with a critical value $f(c)$, where $f(c)$ is a polynomial in $c = w_1 / (w_1 + w_2)$.

Banerjee (1960) and McCullough, Gurland, and Rosenberg (1960) proposed equivalent tests that will be referred to as the BR test. The BR test is similar in form to Cochran's approximate test, with d being compared to an approximate critical value of the form $[w_1 t_{\alpha}^2(f_1) + w_2 t_{\alpha}^2(f_2)]^{\frac{1}{2}} / (w_1 + w_2)$, where $w_i = s_i^2 / n_i$.

Linnik (1966) proved that for the BF situation there is no test with good statistical properties whose level of significance does not depend on the nuisance parameters, the unknown variances. We must content ourselves, therefore, with tests that perform well for a wide range of situations of variance imbalance and sample size imbalance. Some measures of imbalance were considered by Grimes (1979).

Lee and Gurland (1975) proposed a test for the two-sample case with a critical value that depended on the sample variances, the sample sizes, and the nominal significance level. The computation of the critical value involves the solution of a nonlinear minimization problem.

Another large section of literature is devoted to the comparison and evaluation of the proposed solutions to the BF problem. The majority of published results concerning the size and power of the tests are for two-sample comparisons and for limited situations of variance and sample size imbalance. Since the calculation of the exact size and power of these tests is difficult, many studies have been simulation studies. Bennett and Hsu (1961) presented a sampling study of the power functions of the Behrens-Fisher test and the Welch asymptotic test (see Section 2 which describes this test). Their experiment indicated that the BF test showed a smaller level of significance than the Welch test for small sample sizes. The Welch test showed greater power than the BF test over the whole interval of values of $\delta = (\mu_1 - \mu_2) / \sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}$. The complexity of Welch's asymptotic test, however, limits its practical use.

Cochran (1964) published an account of the accuracy of his approximation in the two-sample case, comparing the values of his approximation to Sukhatme's table and an additional table computed by Fisher and Healy (1956). The performance of his approximation appeared to depend upon the significance level, being fairly conservative at the 5% and 1% levels and giving too many significant results at the 10% level.

Golhar (1964) derived the power functions of the tests suggested by Welch and Scheffé for the two-sample case. He calculated the power for a few sample sizes and noncentrality parameters. The calculations involved in evaluating his integral expression for the power function are quite lengthy. Others then began studying the size and power of various tests using numerical integration of Golhar's result. Mehta and Srinivasan (1970) gave several recommendations for the two-sample case. Both the BR and the BF tests had low size and low power for the sample size and variance configurations they studied. Pagurova's test and Welch's asymptotic test had size closer to the supposed nominal level. Both of these tests, however, did not have stable sizes if the larger of the two sample sizes was less than or equal to seven.

Scheffé (1970) discussed several of the proposed solutions to the BF problem. He did not recommend his 1943 solution, and concluded that Welch's approximate solution was the practical one. His rationale was that Welch's solution could be computed with minimal loss of accuracy simply by using the usual Student's *t* tables. For the case of two samples, Wang (1971) compared Welch's approximate test and asymptotic test on the basis of the difference between the actual significance levels of the two tests and the supposed nominal levels of the tests. She used numerical integration techniques to evaluate the probabilities. Her

conclusion, mentioned by Scheffé (1970) prior to its publication, was that in practice one could use the approximate solution with little loss of accuracy.

Lauer and Han (1974) used numerical integration techniques to study the size and power of Cochran's test for the case $k = 2$. If the two sample sizes are equal, Cochran's test is uniformly conservative, that is, for all values of $R = \sigma_1^2 / \sigma_2^2$ the largest actual significance level of Cochran's test is less than or equal to the nominal level. If the sample sizes are not equal, this result still holds except for a small range of values of R . Thus for all practical purposes, they say the test is conservative. They also studied the BR test and found it to be conservative in size. The power of the Cochran test was never smaller than that of the BR test for the situations that they studied.

Lee and Gurland (1975) computed the size and power of several tests for the two-sample case, using numerical integration techniques to evaluate the expression given by Golhar. They reconfirmed the conservativeness of the BR test and the Cochran test. Welch's asymptotic test did the best job in controlling the size, followed closely by the new test proposed by the authors. As stated previously, the complexity of Welch's asymptotic solution limits its practical use. The test of Lee and Gurland is not easily extendible to comparisons of more than two means. Welch's approximate solution showed wider variation than his asymptotic solution, but as Scheffé (1970) stated, it seems a satisfactory practical solution.

Using Monte Carlo techniques, Murphy (1976) also studied the size and power of Welch's approximate solution. He compared its performance to that of the distribution-free Wilcoxon-Mann-Whitney (WMW) test, concluding that the Welch test is acceptable while the WMW test is not. He also discussed some results when samples were drawn from non-normal populations, claiming that if the parent

distribution is reasonably symmetric no serious problem arises. If extreme skewness is present, no test is satisfactory.

Thus in the case of comparing two sample means, the consensus in the literature seems to be approval of Welch's approximate solution and reaffirmation of the conservativeness of Cochran's test. Many standard statistical methods textbooks [see, e.g., Snedecor and Cochran (1967, pages 115-116) and Bliss (1967, pages 216-218)] recommend the use of Cochran's approximate test when variances are unequal. These two tests are then likely candidates for extension to the case of comparisons of more than two means.

In this work, we propose an extension of Cochran's approximation to comparisons of more than two means. These tests and two versions of Welch's test, extended to contrasts of more than two means, are studied and compared in terms of their size and power for specified contrasts under various sample size and variance imbalance situations.

2. FORMULATION OF TEST STATISTICS AND ANALYTIC PROPERTIES

In this section we present analytic descriptions of our four test procedures and analytic results describing their behavior. Recall that we are interested in approximations to the distribution of the sample statistic d for the case $k \geq 2$. We extend Cochran's (1964) approximation for this case by comparing d with an approximate critical value, t'_α , defined as follows

$$t'_\alpha = \frac{\sum_{i=1}^k |c_i| w_i t_\alpha(f_i)}{\sum_{i=1}^k |c_i| w_i} ,$$

where $w_i = s_i^2/n_i$. An alternative value is defined as

$$t''_{\alpha} = \frac{\sum_{i=1}^k c_i^2 w_i t_{\alpha}(f_i)}{\sum_{i=1}^k c_i^2 w_i}.$$

When $f_i = f$ for all k samples in the contrast, both critical values reduce to $t'_{\alpha} = t''_{\alpha} = t_{\alpha}(f)$.

The critical value t'_{α} has one undesirable feature. The statistic d is invariant with respect to changes in the scales of the X_{ij} 's, but t'_{α} is not invariant. We can see this by letting $Z_{kj} = \beta X_{kj}$ where β is a real number which is not equal to zero. If we then divide the value of c_k by β , the value of d does not change. However, the value of t'_{α} does change, and we have

$$t'_{\alpha} = \frac{\sum_{i=1}^{k-1} |c_i| w_i t_{\alpha}(f_i) + \beta |c_k| w_k t_{\alpha}(f_k)}{\sum_{i=1}^{k-1} |c_i| w_i + \beta |c_k| w_k}.$$

The value t''_{α} is invariant to this change in scale.

Two other approximate solutions to the BF problem are also studied. The original idea for the case $k = 2$ put forth by Welch (1938) is that d is distributed approximately as a Student's t , denoted by $d \sim t$. The degrees of freedom for this t are given below in the general case as suggested by Welch (1947). By extension of his proof for the case $k = 2$, we have the following theorem for $k \geq 2$.

Theorem 1. Let $X_{ij} \sim N(\mu_i, \sigma_i^2)$ where the X_{ij} 's are independent. Let
 $d = \frac{\sum_{i=1}^k c_i \bar{X}_i}{\sqrt{\sum_{i=1}^k c_i^2 s_i^2 / n_i}}$ where $\sum_{i=1}^k c_i = 0$ and \bar{X}_i and s_i^2 are the usual
unbiased estimators of μ_i and σ_i^2 , respectively. Then under the null hypothesis
 $\sum_{i=1}^k c_i \mu_i = 0$, d is distributed approximately as a Student's t , with degrees of

freedom, b, given by

$$b = \left[\sum_{i=1}^k \frac{c_i^2 \sigma_i^2}{n_i} \right]^2 / \left[\sum_{i=1}^k \frac{c_i^4 \sigma_i^4}{n_i^2 (n_i - 1)} \right] .$$

In this theorem, when we say that d is distributed approximately as a Student's t , we mean that in deriving the distribution of d the distribution of a weighted sum of Chi-Square random variables is approximated by a Pearsonian type III curve. We choose the parameters of the curve such that the first two moments of the curve agree with the first two moments of the weighted sum.

Since the σ_i^2 are usually unknown, an estimator of b is required. We can replace σ_i^2 in the expression for b by the sample estimator s_i^2 . Letting $\lambda_i = c_i^2/n_i$ and $f_i = n_i - 1$, this yields

$$b1 = \frac{(\sum \lambda_i s_i^2)^2}{(\sum \lambda_i^2 s_i^4 / f_i)} .$$

A second estimator is given by

$$b2 = \frac{(\sum \lambda_i s_i^2)^2 - 2[\sum \lambda_i^2 s_i^4 / (f_i + 2)]}{\sum \lambda_i^2 s_i^4 / (f_i + 2)} .$$

The rationale behind this estimator is that the numerator of $b2$ is an unbiased estimator of $(\sum \lambda_i \sigma_i^2)^2$ and the denominator is an unbiased estimator of $\sum \lambda_i^2 \sigma_i^4 / f_i$. First order Taylor's series approximations for the expected value and variance of a ratio can be used to derive formulas for the expected value and variance of both $b1$ and $b2$. The expressions are lengthy and are given in Grimes (1979; 1981). Both upper and lower bounds on the sizes of $b1$ and $b2$ can be established

as follows:

Theorem 2. $\min_{i=1, \dots, k} (f_i) \leq b1 \leq \sum f_i$ and $\min_{i=1, \dots, k} (f_i) \leq b2 \leq \sum (f_i + 2) - 2$,

where $f_i = n_i - 1$.

Proof. We shall prove $\min_{i=1, \dots, k} (f_i) \leq b1$. Proofs for the other assertions follow similarly. We have

$$b1 = \left(\sum \lambda_i s_i^2 \right)^2 / \left(\sum \lambda_i^2 s_i^4 / f_i \right) \geq \left(\sum \lambda_i s_i^2 \right)^2 / \left(\sum \lambda_i^2 s_i^4 / \min_{i=1, \dots, k} (f_i) \right) \\ \geq \min(f_i) \text{ since } \left(\sum \lambda_i s_i^2 \right)^2 / \left(\sum \lambda_i^2 s_i^4 \right) \geq 1.$$

Thus when all the n_i are equal we can talk about the relative sizes of the tests based on $t''_{\alpha'}$, $b1$, and $b2$.

Theorem 3. If $n_i = n$ for all i then $b2 \geq b1$.

Proof. When sample sizes are equal $b1 = (n-1)(\sum c_i^2 s_i^2)^2 / \sum c_i^4 s_i^4$ and $b2 = (n+1)(\sum c_i^2 s_i^2)^2 / \sum c_i^4 s_i^4 - 2$. Let $x = (\sum c_i^2 s_i^2)^2 / \sum c_i^4 s_i^4$. We want to show that $(n-1)x \leq (n+1)x - 2$ which is equivalent to showing $x \geq 1$. Now $1 \leq (\sum c_i^2 s_i^2)^2 / \sum c_i^4 s_i^4$ if and only if $\sum c_i^4 s_i^4 \leq (\sum c_i^2 s_i^2)^2$. But letting $y_i = c_i^2 s_i^2$ this is equivalent to asking if $\sum y_i^2 \leq (\sum y_i)^2 = \sum y_i^2 + 2 \sum_{i < j} y_i y_j$. This is clearly true since $y_i \geq 0$ for all i .

Let $\alpha_A(x)$ represent the actual size of the test based on the statistic x .

Theorem 4. If $n_i = n$ for all i then $\alpha_A(t''_{\alpha'}) \leq \alpha_A(b1) \leq \alpha_A(b2)$.

Proof. We know that when all samples are of equal size, t''_{α} reduces to a Student's t value based on $n-1$ degrees of freedom. From the previous two theorems we have $b_2 \geq b_1 \geq n-1$. Hence, $\alpha_A(b_2) \geq \alpha_A(b_1) \geq \alpha_A(t''_{\alpha})$.

Another interesting question that can be looked at analytically is how the approximate tests behave for large sample sizes.

Theorem 5. As $\min(n_i)$ for $i = 1, \dots, k$ approaches infinity, t''_{α} approaches Z_{α} , the α percentage point of a standard normal random variable.

Proof. Without loss of generality, assume that we are using values from the positive tail of the t distribution. We then have

$$t''_{\alpha} \leq \left(\sum c_i^2 s_i^2 t_{\alpha}[\max(f_i)]/n_i \right) / \left(\sum c_i^2 s_i^2 / n_i \right) = t_{\alpha}[\max(f_i)] .$$

Thus

$$\lim_{\min(n_i) \rightarrow \infty} t''_{\alpha} \leq \lim_{\min(n_i) \rightarrow \infty} t_{\alpha}[\max(f_i)] = Z_{\alpha} .$$

Also,

$$t''_{\alpha} \geq \left(\sum c_i^2 s_i^2 t_{\alpha}[\min(f_i)]/n_i \right) / \left(\sum c_i^2 s_i^2 / n_i \right) = t_{\alpha}[\min(f_i)] .$$

Hence,

$$\lim_{\min(n_i) \rightarrow \infty} t''_{\alpha} \geq \lim_{\min(n_i) \rightarrow \infty} t_{\alpha}[\min(f_i)] = Z_{\alpha} .$$

Thus,

$$\lim_{\min(n_i) \rightarrow \infty} t''_{\alpha} = Z_{\alpha} .$$

Theorem 6. As $\min(n_i)$ for $i = 1, \dots, k$ approaches infinity, t'_{α} approaches Z_{α} , the α percentage point of a standard normal random variable.

Proof. Proof of this theorem is similar to that of Theorem 5.

Theorem 7. As $\min(n_i)$ for $i = 1, \dots, k$ approaches infinity, $b1$ approaches infinity.

Proof.

$$\begin{aligned} \lim_{\min(n_i) \rightarrow \infty} b1 &= \lim_{\min(n_i) \rightarrow \infty} \left(\sum c_i^2 s_i^2 / n_i \right)^2 / \left(\sum c_i^4 s_i^4 / n_i^2 f_i \right) \\ &\geq \lim_{\min(n_i) \rightarrow \infty} \min(f_i) \left(\sum c_i^2 s_i^2 / n_i \right)^2 / \left(\sum c_i^4 s_i^4 / n_i^2 \right) \\ &\geq \lim_{\min(n_i) \rightarrow \infty} \min(f_i) = \infty. \end{aligned}$$

Theorem 8. As $\min(n_i)$ for $i = 1, \dots, k$ approaches infinity, $b2$ approaches infinity.

Proof. Proof of this theorem is similar to that of Theorem 7.

We can see from theorems 5 through 8 that for large samples all four approximate tests approach a test based on standard normal critical values. This is desirable behavior since it has been established by Mickey and Brown (1966) that in the two-sample case, as the smaller of the two sample sizes approaches infinity, the distribution of d approaches a standard normal.

3. DERIVATION OF ANALYTIC EXPRESSION FOR SIZE AND POWER OF TESTS

To evaluate the behavior of the four tests, we calculated the sizes and powers of each test for a variety of sample size and variance imbalance situations. Our first approach, detailed in Grimes (1979), was to estimate the size

using Monte Carlo simulations. This however was expensive due to the large number of samples needed to maintain even two decimals of accuracy in our estimates. Here we present a more satisfying answer combining an analytic result with numerical analytic techniques. We derive an analytic expression for the size or power of any of the proposed tests, following the approach of Golhar (1964). Recalling the notation given in Section 1 and defining $y = \sum_{i=1}^k c_i \bar{X}_i$, $y \sim N(\sum c_i \mu_i, \sum c_i^2 \sigma_i^2 / n_i)$. Also $(n_i - 1) s_i^2 / \sigma_i^2 \sim \chi_{n_i - 1}^2$, where $\chi_{n_i - 1}^2$ denotes a Chi-Square random variable with $n_i - 1$ degrees of freedom. Let $d = \sum c_i \bar{X}_i / \sqrt{\sum c_i^2 s_i^2 / n_i}$, $\sigma^2 = \sum c_i^2 \sigma_i^2 / n_i$, $\eta = \sum c_i \mu_i$, $\rho = \eta / \sigma$, $f_i = n_i - 1$, and $\lambda_i = c_i^2 / n_i$.

We are considering tests of the null hypothesis $\sum c_i \mu_i = 0$. We base our tests on the sample statistic d . The acceptance region for the hypothesis is $\{d \mid |d| < x^*\}$. To calculate the size or power of any of the multi-sample tests discussed, we need to evaluate

$$\text{Prob} \left\{ \left| y / \sqrt{\sum \lambda_i s_i^2} \right| < x^* \mid \eta, c_i, \sigma_i^2, n_i \text{ for } i = 1, \dots, k \right\}.$$

The value of x^* depends on the test being used and usually involves the s_i^2 . If η is equal to zero, the above probability will equal $1 - \alpha_A$ where α_A is the size of the test based on x^* . If η is not equal to zero, the above probability will equal $1 - \beta_{\alpha_N}$ where β_{α_N} is the power of the test at the supposed nominal level, α_N . To evaluate the above probability, we use numerical quadrature based on the following analysis. First, note that

$$\begin{aligned} & \text{Prob} \left\{ \left| y / \sqrt{\sum \lambda_i s_i^2} \right| < x^* \mid \eta, c_i, \sigma_i^2, n_i \text{ for } i = 1, \dots, k \right\} \\ &= \text{Prob} \left\{ |y| < x^* \sqrt{\sum \lambda_i s_i^2} \mid \eta, c_i, \sigma_i^2, n_i \text{ for } i = 1, \dots, k \right\}. \end{aligned}$$

The \bar{X}_i are independent of the s_j^2 for all i and j , and we know the distributions of \bar{X}_i and s_j^2 ; hence,

$$P = \text{Prob} \left\{ |y| < x^* \sqrt{\sum \lambda_i s_i^2} \mid \eta, c_i, \sigma_i^2, n_i \text{ for } i = 1, \dots, k \right\}$$

$$= B \int_0^\infty \dots \int_0^\infty \int_{-x^* \sqrt{\sum \lambda_i \sigma_i^2 V_i}}^{x^* \sqrt{\sum \lambda_i \sigma_i^2 V_i}} \prod_{i=1}^k V_i^{\{f_i/2\}-1} \exp \left\{ - \sum f_i V_i / 2 \right\}$$

$$\cdot \sigma^{-1} \exp \left\{ - [(y-\eta)/\sqrt{2}\sigma]^2 \right\} dy \, dV_i, \quad i = 1, \dots, k,$$

where

$$B = \left(2^{\{\sum f_i/2\}} \sqrt{2\pi} \prod_{i=1}^k \Gamma(f_i/2) \right)^{-1} \prod_{i=1}^k f_i^{\{f_i/2\}}$$

and

$$V_i = s_i^2 / \sigma_i^2.$$

Now, letting $U = (y-\eta)/\sigma$, we have

$$P = B \int_0^\infty \dots \int_0^\infty \prod_{i=1}^k V_i^{\{f_i/2\}-1} \exp \left\{ - \sum f_i V_i / 2 \right\} \int_{-a-\rho}^{a-\rho} \exp \{-U^2/2\} dU \, dV_i,$$

$$i = 1, \dots, k,$$

where $a = x^* \sqrt{\sum \lambda_i \sigma_i^2 V_i} / \sigma$. Thus to calculate the power or the size of any of our tests, we must evaluate this multidimensional integral. Table 3.1 gives the expression for x^* for each of the tests we are considering. Since $\int_{-\infty}^x (2\pi)^{-\frac{1}{2}} \exp\{-U^2/2\} du = \Phi(x)$, the integral to be evaluated reduces to

$$C \int_0^\infty \int_0^\infty \dots \int_0^\infty \exp \left\{ - \sum_{i=1}^k f_i V_i / 2 \right\} \prod_{i=1}^k V_i^{\{f_i/2\}-1} [\Phi(a-\rho) - \Phi(-a-\rho)] dV_1 \dots dV_k,$$

where

TABLE 3.1

Expressions for x in Evaluation of Power or Size of Tests

Based on t_{b1} , t_{b2} , t'_{α} , and t''_{α} . Note $V_i = s_i^2/\sigma_i^2$.

Test based on	x^* (critical value)
t_{b1}	$t_{\alpha} \left[\left(\sum \lambda_i \sigma_i^2 V_i \right)^2 / \sum \lambda_i^2 \sigma_i^4 V_i^2 / f_i \right]$
t_{b2}	$t_{\alpha} \left[\left[\left(\sum \lambda_i \sigma_i^2 V_i \right)^2 / \sum \lambda_i^2 \sigma_i^4 V_i^2 / (f_i + 2) \right] - 2 \right]$
t'_{α}	$\left(\sum c_i \sigma_i^2 V_i t_{\alpha}(f_i) / n_i \right) / \left(\sum c_i \sigma_i^2 V_i / n_i \right)$
t''_{α}	$\left(\sum \lambda_i \sigma_i^2 V_i t_{\alpha}(f_i) \right) / \left(\sum \lambda_i \sigma_i^2 V_i \right)$

$t_{\alpha}(x)$ is the value z such that $\text{Prob}\{|t| < z\} = 1 - \alpha$, where t is a Student's t random variable based on x degrees of freedom.

$$c = \left[2^{\{\sum f_i/2\}} \prod_{i=1}^k \Gamma(f_i/2) \right]^{-1} \prod_{i=1}^k f_i^{f_i/2} .$$

Remember that x^* depends on V_1, \dots, V_k . Thus letting $g(V_1, \dots, V_k) = \Phi(a-\rho) - \Phi(-a-\rho)$, the integral to be evaluated is of the form

$$c \int_0^\infty \dots \int_0^\infty \exp \left\{ - \sum_{i=1}^k f_i V_i / 2 \right\} \prod_{i=1}^k V_i^{\{f_i/2\}-1} g(V_1, \dots, V_k) dV_1 \dots dV_k ,$$

where

$$0 \leq g(V_1, \dots, V_k) \leq 1 \text{ for all } V_i, \quad i = 1, \dots, k .$$

The choice of a numerical technique for evaluating the multidimensional integral depends upon the dimensionality of the problem. For low dimensional problems, i.e., $k = 2, 3$, or 4 , a quadrature rule is known to be more effective than a Monte Carlo integration technique (Bell and Glaz, 1980). Hence a quadrature rule was used for the low dimensional problems, and a Monte Carlo integration technique was used for the higher dimensional problems. The details are presented in Grimes (1979;1981).

4. PRESENTATION OF SIMULATION RESULTS

4.1. SELECTION OF EXPERIMENTAL FACTORS

We present results on the size and power of the four tests for comparisons involving three, four, and eight means. Since the size and power of tests in the BF situation depend upon the population variances, we must examine the

behavior of our tests for a wide range of variance imbalance situations. Hence we are in the role of an experimenter who must first decide which factors are most influential on the response of interest and then what range of values of these factors will give a good representation of the response.

In this study we varied four factors: 1) the value of $\Sigma c_i \mu_i$, 2) population variances, 3: sample sizes, and 4) the nominal significance level. We express the value of the contrast of population means as a multiple of the standard error, i.e., we set $\eta = \rho\sigma$. This is equivalent to looking at various values of the noncentrality factor ρ .

The study was divided into two phases. First we considered comparisons involving $k = 3$ samples. We used the contrast $(1,1,-2)$, i.e., a comparison of the form $\bar{X}_1 + \bar{X}_2 - 2\bar{X}_3$. Table 4.1 gives the values used for ρ , α_N , and σ_i^2 , $i = 1,2,3$ and the variance and sample size imbalance situations considered. The size of each test was not evaluated for all $5 \times 4 \times 5 \times 4 = 400$ combinations. Values of the first four levels of two of the factors and all four levels of the other two would have entailed 256 evaluations. We selected 16 of the 256 combinations for initial information by using an orthogonal main effects plan. The results are presented as the first sixteen entries in Table 4.2, as well as the details of sample sizes and variances for each of the combinations. Based on the information gained from these, we selected other combinations to include the fifth level for two of the factors. The vector of population means is $\underline{\mu}$, $\underline{\sigma}^2$ is the vector of population variances, and \underline{n} is the vector of sample sizes.

We used the results of phase one to guide our selection of factor values for the second phase of our study in which we looked at contrasts involving four and eight means. To determine if the tests behaved differently for different

TABLE 4.1

Factors Selected for Preliminary Evaluation of Tests

Level	A = ρ	B = Sample Size and Variance Situations	C = α_N	D = σ^2
0	0	σ_1^2/n_1 equals a constant	.10	(1, 2 $\frac{1}{3}$, 3)
1	$\frac{1}{2}$	$n_1\sigma_1^2$ equals a constant	.05	(1, 3, 5)
2	1	small equal n_1 , σ_1^2 different	.025	(1, 5, 7)
3	2	large equal n_1 , σ_1^2 different	.01	(1, 1, 7)
4		small unequal n_1 , σ_1^2 different		(1, 1, 3)

TABLE 4.2

Actual Size or Power of Tests Based on t''_{α} , t'_{α} , t_{b1} , and t_{b2}
for Given Combinations of Factors from Table 4.1

Combination	$\Sigma c_i \mu_i$	σ^2	n	α_N	Tests			
					t''_{α}	t'_{α}	b1	b2
1) 0000	0	(1, 2 $\frac{1}{3}$, 3)	(9, 21, 27)	.10	.0895	.0871	.0997	.1001
2) 0111	0	(1, 3, 5)	(45, 15, 9)	.05	.0460	.0474	.0518	.0531
3) 0222	0	(1, 5, 7)	(5, 5, 5)	.025	.0111		.0273	.0318
4) 0333	0	(1, 1, 7)	(21, 21, 21)	.01	.0091		.0102	.0103
5) 1013	$\frac{1}{2}\sigma$	(1, 1, 7)	(5, 5, 35)	.05	.0497	.0399	.0760	.0778
6) 1102	$\frac{1}{2}\sigma$	(1, 5, 7)	(35, 7, 5)	.10	.1161	.1194	.1360	.1424
7) 1231	$\frac{1}{2}\sigma$	(1, 3, 5)	(5, 5, 5)	.01	.0046		.0192	.0239
8) 1320	$\frac{1}{2}\sigma$	(1, 2 $\frac{1}{3}$, 3)	(21, 21, 21)	.025	.0367		.0426	.0430
9) 2021	σ	(1, 3, 5)	(9, 27, 45)	.025	.0871	.0815	.1043	.1050
10) 2130	σ	(1, 2 $\frac{1}{3}$, 3)	(21, 9, 7)	.01	.0264	.0287	.0455	.0498
11) 2203	σ	(1, 1, 7)	(5, 5, 5)	.10	.2037		.2263	.2335
12) 2312	σ	(1, 5, 7)	(21, 21, 21)	.05	.1530		.1623	.1630
13) 3032	2 σ	(1, 5, 7)	(9, 45, 63)	.01	.2281	.2123	.2708	.2721
14) 3123	2 σ	(1, 1, 7)	(35, 35, 5)	.025	.2083	.2115	.2151	.2197
15) 3301	2 σ	(1, 3, 5)	(21, 21, 21)	.10	.6118		.6206	.6211
16) 3210	2 σ	(1, 2 $\frac{1}{3}$, 3)	(5, 5, 5)	.05	.2990		.3970	.4184
17) 1221	$\frac{1}{2}\sigma$	(1, 3, 5)	(5, 5, 5)	.025	.0181		.0412	.0474
18) 1331	$\frac{1}{2}\sigma$	(1, 2 $\frac{1}{3}$, 3)	(21, 21, 21)	.01	.0150		.0192	.0195
19) 0231	0	(1, 3, 5)	(5, 5, 5)	.01	.0026		.0124	.0156
20) 0331	0	(1, 2 $\frac{1}{3}$, 3)	(21, 21, 21)	.01	.0076		.0102	.0103
21) 0212	0	(1, 5, 7)	(5, 5, 5)	.05	.0301		.0520	.0577
22) 0111	0	(1, 3, 5)	(15, 5, 3)	.05	.0362	.0422	.0671	.0818
23) 0000	0	(1, 2 $\frac{1}{3}$, 3)	(3, 7, 9)	.10	.0621	.0552	.0963	.1008
24) 0213	0	(1, 1, 7)	(5, 5, 5)	.05	.0395		.0550	.0595
25) 0214	0	(1, 1, 3)	(5, 5, 5)	.05	.0318		.0541	.0597
26) 0413	0	(1, 1, 7)	(3, 5, 7)	.05	.0289	.0241	.0500	.0534
27) 0414	0	(1, 1, 3)	(3, 5, 7)	.05	.0210	.0168	.0472	.0521
28) 0414	0	(1, 1, 3)	(7, 5, 3)	.05	.0379	.0445	.0713	.0863
29) 0413	0	(1, 1, 7)	(7, 5, 3)	.05	.0439	.0494	.0688	.0824
30) 0410	0	(1, 2 $\frac{1}{3}$, 3)	(3, 7, 9)	.05	.0213	.0174	.0471	.0509
31) 0430	0	(1, 2 $\frac{1}{3}$, 3)	(3, 7, 9)	.01	.0012	.0008	.0089	.0106

The contrast used is (1,1,-2), i.e., a contrast of the form $\mu_1 + \mu_2 - 2\mu_3$.

contrast types we also used two different types of contrasts. The variances, sample sizes, and contrasts are given in Table 4.3. For example, looking at row 1 of the variance structure matrix, the entries represent $\sigma_1^2 = 1$, $\sigma_2^2 = 1.5$, $\sigma_3^2 = 2$, and $\sigma_4^2 = 4$. Looking at the contrast matrix, rows 1 and 2 represent the coefficients of orthogonal polynomial contrasts and rows 3 and 4 represent the coefficients of Hadamard contrasts.

In Table 4.4 the sizes of the tests as calculated by using the Monte Carlo integration routine are presented for the combinations given in the second column. For example, combination (1,1,1) means (referring to Table 4.3) that the contrast is of the form $-1\bar{X}_1 + 3\bar{X}_2 - 3\bar{X}_3 + \bar{X}_4$ where independent samples of size 5 are taken from populations with variances of 1, 1.5, 2, and 4, respectively. In Table 4.5, the power of each of the four tests is given against the alternative $\sum c_i \mu_i = \sigma$, i.e., $\rho = 1$ for the specified combinations.

We also looked at comparisons involving eight samples. Table 4.6 gives the contrasts, variance structures, and sample sizes from which we selected some combinations. Tables 4.7 and 4.8 present the sizes and powers, respectively, for our selections.

4.2. DISCUSSION OF SIMULATION RESULTS

We will first discuss Table 4.2 in which results are presented for comparisons involving three means. Recall that combinations 1 through 16 represent our preliminary study. When looking at these combinations, it was clear that the tests were becoming asymptotically equivalent when each sample size was 21 or larger. Hence when selecting additional combinations, we focused our attention on small sample sizes.

TABLE 4.3

Contrasts, Variance Structures, and Sample Sizes
for the Four Populations Case

Contrasts				
1	-1	3	-3	1
2	-3	-1	1	3
3	1	-1	-1	1
4	1	1	-1	-1

Variance Structures				
1	1.0	1.5	2.0	4.0
2	1.0	2.0	4.0	8.0
3	1.0	1.0	1.0	8.0

Sample Sizes				
1	5	5	5	5
2	5	5	5	15
3	15	15	15	5
4	21	21	21	21
5	5	7	9	11
6	11	9	7	5

TABLE 4.4

Actual Sizes of Tests Based on t''_{α} , t'_{α} , t_{b1} , and t_{b2} for $\alpha_N = .05$ and
Given Contrasts, C, Variance Structures, VS, and Sample Sizes, SS

Number	(C, VS, SS)	$\alpha_A(t''_{\alpha})$	$\alpha_A(t'_{\alpha})$	$\alpha_A(t_{b1})$	$\alpha_A(t_{b2})$
1	(1,1,1)	.018		.046	.052
2	(1,2,1)	.019		.047	.053
3	(1,3,1)	.016		.045	.051
4	(1,3,2)	.021	.026	.048	.053
5	(1,3,3)	.031	.024	.052	.056
6	(2,1,4)	.044		.050	.050
7	(2,2,4)	.046		.050	.051
8	(2,3,2)	.030	.027	.049	.050
9	(2,3,4)	.047		.050	.051
10	(3,1,2)	.018		.047	.051
11	(3,2,2)	.020		.048	.052
12	(3,2,3)	.037		.057	.062
13	(3,3,4)	.044		.050	.051
14	(4,1,1)	.016		.046	.052
15	(1,1,5)	.029	.028	.049	.051
16	(1,1,6)	.026	.023	.048	.051
17	(1,2,5)	.029	.030	.048	.050
18	(1,2,6)	.026	.023	.048	.051

TABLE 4.5

Power of Tests Based on t''_{α} , t'_{α} , t_{b1} , and t_{b2} for $\alpha_N = .05$ and Given
Contrasts, C, Variance Structures, VS, and Sample Sizes, SS
Against the Alternative $\sum c_i \mu_i = \sigma$

Number	(C, VS, SS)	$\beta_{.05}(t''_{\alpha})$	$\beta_{.05}(t'_{\alpha})$	$\beta_{.05}(t_{b1})$	$\beta_{.05}(t_{b2})$
1	(1,1,6)	.097	.088	.151	.157
2	(1,2,6)	.097	.088	.150	.157
3	(1,2,5)	.109	.109	.154	.159

TABLE 4.6

Contrasts, Variance Structures, and Sample Sizes
for the Eight Samples Case

Contrasts

1	[1	1	1	1	-1	-1	-1	-1]
2	[15	17	-23	7	-7	23	-17	-15]

Variance Structures

1	[1	1	4	4	7	7	10	10]
2	[1	1.5	2	3	4	6	7	9]

Sample Sizes

1	[7	7	9	9	11	11	13	13]
2	[5	5	7	7	9	9	11	11]
3	[11	11	9	9	7	7	5	5]

TABLE 4.7

Actual Size of Tests Based on t''_{α} , t'_{α} , t_{b1} , and t_{b2} for Given Contrast,
Variance Structure, and Sample Size Combination¹, and
Nominal Significance Level of 0.05

Number	(C, VS, SS) ²	$\alpha_A(t''_{\alpha})$	$\alpha_A(t'_{\alpha})$	$\alpha_A(t_{b1})$	$\alpha_A(t_{b2})$
1	(1,1,1)	.028		.049	.050
2	(2,2,2)	.023	.023	.049	.050
3	(2,2,3)	.018	.018	.048	.051

¹ The contrasts, variance structures, and sample sizes are those given in Table 4.6. No entry for t'_{α} appears for the first combination because for Hadamard contrasts $t''_{\alpha} = t'_{\alpha}$.

² C = contrast, VS = variance structure, SS = sample size.

TABLE 4.8

Power of Tests Based on t''_{α} , t'_{α} , t_{b1} , and t_{b2} for Given Contrast, Variance Structure, and Sample Size Combination¹, and Nominal Significance Level of 0.05 Against the Alternative $\Sigma c_i \mu_i = \sigma$

Number	(C, VS, SS) ²	$\beta_{.05}(t''_{\alpha})$	$\beta_{.05}(t'_{\alpha})$	$\beta_{.05}(t_{b1})$	$\beta_{.05}(t_{b2})$
1	(2, 2, 2)	.097	.097	.162	.165
2	(2, 2, 3)	.076	.075	.152	.160

¹ The contrasts, variance structures, and sample sizes are those given in Table 4.6.

² C = contrast, VS = variance structure, SS = sample size.

Looking only at the cases where $\sum c_i \mu_i = 0$, we can see that the tests can be ordered in terms of their conservativeness. We have $\alpha_A(t''_{\alpha}) < \alpha_A(t_{b1}) < \alpha_A(t_{b2})$. (Note that since the tests based on t''_{α} and t'_{α} behave much the same, we limit our discussion to the invariant test based on t''_{α} .) In all cases the test based on t''_{α} is conservative. For small sample sizes the ratio between the actual significance level and the nominal level for t''_{α} appears to vary with the nominal level. For example, at $\alpha_N = .01$, $\alpha_A(t''_{\alpha})$ took on values .0026 and .0012, approximately one-fourth to one-eighth of the supposed nominal level. At $\alpha_N = 0.05$, $\alpha_A(t''_{\alpha})$ equals approximately two-thirds of α_N .

The test based on t_{b2} is liberal in every case. The actual significance level of the test based on t_{b1} fluctuates around the nominal level. The amount of difference between the actual and nominal significance levels is, however, independent of the nominal level. When comparing the absolute value of the difference between the actual significance level and the nominal level, we see that in most cases the value is smaller for t_{b1} and t_{b2} than for t''_{α} . The exception occurs when smaller sample sizes are taken from the more variable populations. Here if we use the test based on t_{b2} , our actual significance level may be almost double the supposed nominal level. The test based on t_{b1} performs better, but it is still liberal.

We also evaluated the power of the various tests. Since the tests are all operating at different actual significance levels, it is not strictly fair to compare their powers. At a given nominal level, a test operating at a higher significance level would appear to have greater power than another test operating at a lower significance level even if the power curves are actually the same. However, since the practitioner is essentially forced to deal with nominal significance levels, we will discuss results from this point of view. Again

looking at Table 4.2, we see that $\beta_{\alpha_N}(t_{b2}) > \beta_{\alpha_N}(t_{b1}) > \beta_{\alpha_N}(t''_{\alpha})$ for all combinations. The fact that the test based on t''_{α} is less powerful is not surprising due to the conservativeness of the test. It should also be noted that in most cases, the power of the test based on t_{b1} differs from the power of the test based on t_{b2} only in the third or fourth decimal place.

To summarize this study we found:

1. The test based on t''_{α} is conservative, especially for small sample sizes and small nominal significance levels,
2. the tests based on $b1$ and $b2$ are liberal,
3. the test based on $b1$ has α_A close to α_N except when the smaller samples are drawn from the more variable populations, and
4. we see the asymptotic equality of the tests when each sample size is 21 or greater.

We next wanted to study situations where four or more means are in the contrast. The contrasts, variance structures, and sample sizes considered are presented in Table 4.3. To examine more situations of sample size and variance imbalance, all the tests were compared at a nominal significance level of .05. In Tables 4.4 and 4.5, the results for contrasts of four means are presented. The results seen above for the use of 3 populations again hold, except that the size associated with t_{b1} tends to be less than the nominal level in most cases. We wished to see if the tests behaved differently for different contrast types. Comparing the results for contrasts 1 and 2 with the results for contrasts 3 and 4, we see no change in the behavior of the tests.

Finally we looked at comparisons involving eight samples (Tables 4.6 - 4.8). We took the maximum sample size to be 13 as we know that for large sample sizes the tests are almost indistinguishable with respect to size and power. We also

used an orthogonal polynomial contrast with large coefficients to see if the magnitude of the coefficients would affect the behavior of the tests. Again we see the same results as for comparisons involving fewer means. The magnitude of the coefficients does not influence the behavior.

5. SUMMARY

Based on the results of our investigation we can summarize our conclusions and recommendations as follows:

1. The type of contrast or magnitude of coefficients in the contrast had no effect on the behavior of the tests;
2. for sample sizes ≥ 15 or greater, the tests begin to demonstrate their asymptotic equality;
3. the test based on t''_{α} is conservative, especially when sample sizes are small;
4. the test based on t_{b1} is preferable to the test based on t_{b2} , since the test based on t_{b1} is less liberal and has comparable power;
5. the test based on t''_{α} should be used if false rejection of the null hypothesis is costly;
6. the test based on $b1$ should be used if detection of differences is more important; and
7. the programs developed for evaluation of the size and power can be used for other values of x^* than the ones we have currently studied.

REFERENCES

- Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. Biometrika 35, 88-96.
- Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. Biometrika 36, 290-296.
- Banerjee, S. K. (1960). Approximate confidence interval for linear functions of means of k populations when the populations' variances are not equal. Sankhyā 22, 357-358.
- Behrens, W. U. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. Landwirtschaftliche Jahrbücher 68, 807-837.
- Bell, J., and Glaz, H. (1980). Private communication.
- Bennett, B. M., and Hsu, P. (1961). Sampling studies on the Behrens-Fisher problem. Metrika 4, 89-105.
- Bliss, C. I. (1967). Statistics in Biology. McGraw-Hill, Inc., New York.
- Cochran, W. G. (1964). Approximate significance levels of the Behrens-Fisher test. Biometrics 20, 191-195.
- Federer, W. T. (1955). Experimental Design. Macmillan, New York.
- Fisher, R. A. (1935). The fiducial argument in statistical inference. Annals of Eugenics 6, 391-398.
- Fisher, R. A. (1941). The asymptotic approach to Behrens' integral, with further tables for the d test of significance. Annals of Eugenics 11, 141-172.
- Fisher, R. A., and Healy, M. J. R. (1956). New tables of Behrens' test of significance. Journal of the Royal Statistical Society, Series B 18, 212-216.
- Golhar, M. B. (1964). On the comparison of two means from normal populations with unknown variances. Indian Society of Agricultural Statistics 16, 62-71.

- Grimes, B. A. (1979). Cochran-like and Welch-like approximate solutions to the problem of comparison of means from two or more populations with unequal variances. Master's Thesis, Cornell University, Ithaca, New York.
- Grimes, B. A. (1981). On the comparison of means from populations with unequal variances. Ph.D. Thesis, Cornell University, Ithaca, New York.
- Lauer, G. N., and Han, C. P. (1974). Power of Cochran's test in the Behrens-Fisher problem. Technometrics 16, 545-549.
- Lee, A. F. S., and Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. Journal of the American Statistical Association 70, 933-941.
- Linnik, Y. V. (1966). Latest investigations on Behrens-Fisher problem. Sankhyā 28, 15-24.
- McCullough, R. S., Gurland, J., and Rosenberg, L. (1960). Small sample behaviour of certain tests of the hypothesis of equal means under variance heterogeneity. Biometrika 47, 345-353.
- Mehta, J. S., and Srinivasan, R. (1970). On the Behrens-Fisher problem. Biometrika 57, 649-655.
- Mickey, M. R., and Brown, M. B. (1966). Bounds on the distribution function of the Behrens-Fisher statistic. The Annals of Mathematical Statistics 37, 639-642.
- Murphy, B. P. (1976). Comparison of some two-sample means tests by simulation. Communications in Statistics, Simulation and Computation B5, 23-32.
- Pagurova, V. I. (1968). On a comparison of means of two normal samples. Theory of Probability and Its Applications 13, 527-534.
- Scheffé, H. (1943). On solutions of the Behrens-Fisher problem, based on the t-distribution. The Annals of Mathematical Statistics 14, 35-44.
- Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problems. Journal of the American Statistical Association 65, 1501-1508.
- Snedecor, G. W. (1946). Statistical Methods, 4th edition. Iowa State University Press, Ames, Iowa.

- Snedecor, G. W., and Cochran, W. G. (1967). Statistical Methods, 6th edition. Iowa State University Press, Ames, Iowa.
- Sukhatme, P. V. (1938). On Fisher and Behrens' test of significance for the difference in means of two normal samples. Sankhyā 4, 39-48.
- Wald, A. (1955). Testing the difference between means of two normal populations with unknown standard deviations. Selected Papers in Probability and Statistics 1, 669-695.
- Wang, Y. Y. (1971). Probabilities of the Type 1 errors of the Welch tests for the Behrens-Fisher problem. Journal of the American Statistical Association 66, 605-608.
- Welch, B. L. (1936). The specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem. Journal of the Royal Statistical Society, Supplement B, 29-48.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika 29, 350-362.
- Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34, 28-35.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika 38, 330-336.